RCJ Journal of the Academy of Public Health

A Proof-of-Concept Metadata Catalogue and Online Portal of Earth Observation Datasets for Health Research in Exposomics

Keumseok Koh, Maged N. Kamel Boulos, Gang Zheng, Hongsheng Zhang, Muralikrishna V. Iyyanki, Bosco Bwambale, Ashraf Dewan

Published: Jan 30, 2025 + DOI: https://doi.org/10.70542/rcj-japh-art-1wjzmn4

Abstract

This article describes research carried out during 2023 and the first quarter of 2024 under an International Society for Photogrammetry and Remote Sensing (ISPRS)-funded project to develop and disseminate a metadata catalogue of Earth observation data sources/products and types that are relevant to human health research in exposomics, as a free service to interested researchers worldwide. The proof-of-concept catalogue was informed by input from existing research literature on the subject (desk research), as well as online communications with, and relevant research publications collected from, a small panel (n = 5) of select experts from the academia in three countries (China, UK and USA). It has 90 metadata records of relevant Earth observation datasets (n = 40) and associated health-focused research publications (n = 50). The project's online portal offers a searchable version of the catalogue featuring a number of search modes and filtering options. It is hoped future, more comprehensive versions of this service will enable more researchers and studies to discover and use remote sensing data about population-level exposures to disease determinants (exposomic determinants of disease) in combination with other relevant data to reveal fresh insights that could improve our understanding of relevant diseases, and hence contribute to the development of better-optimized prevention and management plans to tackle them.

Keywords: exposomics; human exposome; Earth observation data; remote sensing; data sources; metadata catalogue; public health informatics

Disclosures, Funding & Conflicts of Interest:
The proof-of-concept catalogue data covered in this paper are publicly and freely available from doi: 10.5281/zenodo.8186444 and https://geoexposomics.org/
This research was funded by the International Society for Photogrammetry and Remote Sensing (ISPRS) under ISPRS Scientific Initiatives 2023 (SI2023 Awards).
The authors declare no conflict of interest.

Affiliations:
Keumseok Koh, PhD, The University of Hong Kong, Hong Kong, China
Maged N. Kamel Boulos, PhD, University of Lisbon, Lisbon, Portugal
Gang Zheng, PhD, The University of Hong Kong, Hong Kong, China
Hongsheng Zhang, PhD, The University of Hong Kong, Hong Kong, China
Muralikrishna V. Iyyanki, PhD, JN Technological University Hyderabad, Hyderabad, India
Bosco Bwambale, PhD, Mountains of The Moon University, Fort-Portal, Uganda
Ashraf Dewan, PhD, School of Earth & Planetary Sciences, Curtin University, Perth, Australia

Correspondence:
mnkboulos@ieee.org

Submitted 11/12/2024

Published by Journal of the Academy of Public Health. This is an Open Access article under the CC BY 4.0 license.

1 - Introduction

1.1 Earth Observation Data

Earth observation data include remote sensing (RS) data from satellites, aerial platforms, drones, and seaborne and ground-based sensors, among others. Typically obtained by capturing and analyzing the energy reflected or emitted from the Earth's surface, atmosphere, or other objects of interest,[1] RS data provide valuable insights into various aspects of the Earth's environment, such as air pollutants, land cover, green space, topography, nocturnal outdoor light pollution, and noise pollution, among others. Moreover, advances in wearable sensors and mobile technologies are further expanding the applications of RS. RS data, combined with other human population data collected at multiple scales are being successfully applied to various real-life problems and challenges to better manage them.[2]

1.2 Human Exposomics

Predisposition and development of various diseases, communicable, such as malaria and dengue fever, and non-communicable, such as diabetes, cardiovascular diseases, various types of cancer and mental health problems, involve a complex interplay between genetic factors (the genome) and environmental and lifestyle parameters that populations are exposed to (the exposome).[3] When combined with other relevant data, RS data from various sources/products and of different types can help us better map and investigate the latter (exposomic determinants of disease/population-level exposures).[4,5,6,7,8] However, these Earth observation data sources and types remain a largely untapped resource for many researchers in the fields of population health and medicine, who are not familiar with the potential and value of these data and the unique insights that can be revealed by using them.

This article describes the authors' work on an International Society for Photogrammetry and Remote Sensing (ISPRS)-funded project entitled 'geospatial database for exposomics (development of an online portal and metadata catalogue of Earth observation data types, sources and products for human health research in exposomics)'. This project is funded under ISPRS Scientific Initiatives 2023 (SI2023 Awards),[9] and involves proof concepting and disseminating a much-needed metadata catalogue of Earth observation data sources/products and types that are relevant to human health research as a free service to interested researchers worldwide. Even though the various uses of RS data in large population health studies are not a new thing as such (existing studies actually form the basis of our project), our rationale is to further highlight and facilitate such uses and open them up to many more researchers and studies in the fields of population health and medicine.

2 - Materials and Methods

2.1 Catalogue Curation: External Expert Involvement and Metadata Elements

The project's proof-of-concept catalogue was informed by a rather extensive input from existing research literature on the subject (desk research) as well as online communications with, and relevant research publications collected from, a small panel (n = 5) of select 'stakeholder representatives' from the academia (backgrounds in health/human geography and/or geomatics) in three countries (China, UK and USA) to guarantee some basic form of user engagement and involvement in the project from its very early stages, which is always highly desirable.

This small panel of five academics was a convenience sample, given the project team's human resource, time and budget constraints and the fact that in this project we are only proof concepting the catalogue and it was never our goal to comprehensively catalogue every RS dataset and application in health under the sun. Ten academics were initially invited via email, but only five of them responded. All invited academics were external experts (i.e., not members of the project team) but well-known to us for their leading research into RS applications in health.

Online communications with panel members took place between February and June 2023. The five stakeholder representatives provided us with valuable insights and intelligence about relevant remote sensing applications and available products in public health, in addition to important feedback on our catalogue during its development over two iterations in 2023, which helped us improve and refine it to best serve their needs and those of the wider research community they represent.

The research team compiled the catalogue based on all collected input in the form of an Excel workbook.[10] The

Published by Journal of the Academy of Public Health. This is an Open Access article under the CC BY 4.0 license.

catalogue currently has 90 metadata records of relevant Earth observation datasets (n = 40) and associated health-focused research publications (n = 50). Dataset and research publication records are stored in two separate but fully cross-referenced Excel worksheets/database tables. For each catalogued record, the following metadata elements were compiled: dataset name, provider, available year(s)/update frequency, cost (license), covered geographical areas, resolution used, URL, related research publications, and the health conditions/application areas that featured in those publications.

2.2 Web Conversion

Our Web development team then converted the Excel workbook into an online database and developed the project's online portal in MySQL, an open-source relational database management system; Python, a high-level, general-purpose programming language; and Django, an open source, Python-based Web framework. The portal's public user interface was designed with both desktop and mobile devices in mind. The portal also features an access-protected 'Catalogue Management Interface' where the catalogue curator can modify existing catalogue records or add new ones as necessary.

3 - Results

3.1 Catalogue Summary Statistics: By Publications (n = 50)

3.1.1 Journal Categories

Approximately 40% (n = 21) of curated articles (health-focused research publications) in our proof-of-concept catalogue were published in geography-themed journals, followed by public health (n = 13), environment (n = 9), and science-themed (n = 7) journals.

3.1.2 Study Areas

The majority of curated studies in our proof-of-concept catalogue are global studies (n = 30). The remaining studies cover North America/USA (n = 9), Europe/UK (n = 7), Asia (n = 2), Africa (n = 1), and the Global Coastal Zone (n = 1). Readers can refer to [10] for detailed information about the curated publications and their corresponding study areas.

3.1.3 Study Topics

Table 1 summarizes the research topics of the curated publications in the proof-of-concept catalogue. We broadly classified the topics into three sub-categories or themes—human activities, public health, and the environment. Almost half of the publications (n = 24) investigated various themes of human activities, followed by public health (n = 14), and the environment (n = 12). Readers can refer to [10] for detailed information about which curated publication(s) is/are covering which topic, with full citation details.

Published by Journal of the Academy of Public Health. This is an Open Access article under the CC BY 4.0 license.

Table 1. Study topics.

Application Areas	Study Topics	Number of Works
Human Activity (including anthropogenic activity affecting the environment)	Children Physical Activity (n = 4), Community-Level Livability (n = 1), Humanitarian Emergencies (n = 1), Land Use Mix Index (n = 1), Landscape Disturbance (n = 1), Nighttime Lights (n = 2), Population Exposure (n = 1), Residential Environments Greenness (n = 1), Transportation Network (n = 1), Urban Expansion (n = 2), Urban Foraging (n = 1), Urban Green Spaces (n = 2), Urban Sustainability (n = 2), Urban Temperatures/Thermal Environment (n = 3), Urban-associated Salmonellae (n = 1)	24
Public Health	Anxiety Symptoms during the COVID-19 Pandemic (n = 1), Hemorrhagic Fever (n = 4), Hand, Foot, and Mouth Disease (HFMD) Outbreak (n = 1), Human Health and well-being (n = 1), Human Rabies (n = 1), Influenza Virus (n = 1), Malaria Prevention (n = 1), Rural Mental Health (n = 1), Salmonella Serovars and Strains (n = 1), Sociodemographic Environments and Obesity (n = 1), West Nile Virus (n = 1)	14
Environment	Air Pollution (n = 2), Grassland (Prairie) Ecosystems (n = 2), Mixed Grasslands (n = 1), Mountain Green Cover Index (n = 1), Ocean Harmful Algal Blooms (n = 4), Rural Water (n = 2)	12
Total		50

3.1.4 Study Publication Years

Half of the articles (n = 25) were published before 2018. Remarkably, 25% of the articles (n = 12) were published in 2020 alone.

Published by Journal of the Academy of Public Health. This is an Open Access article under the CC BY 4.0 license.

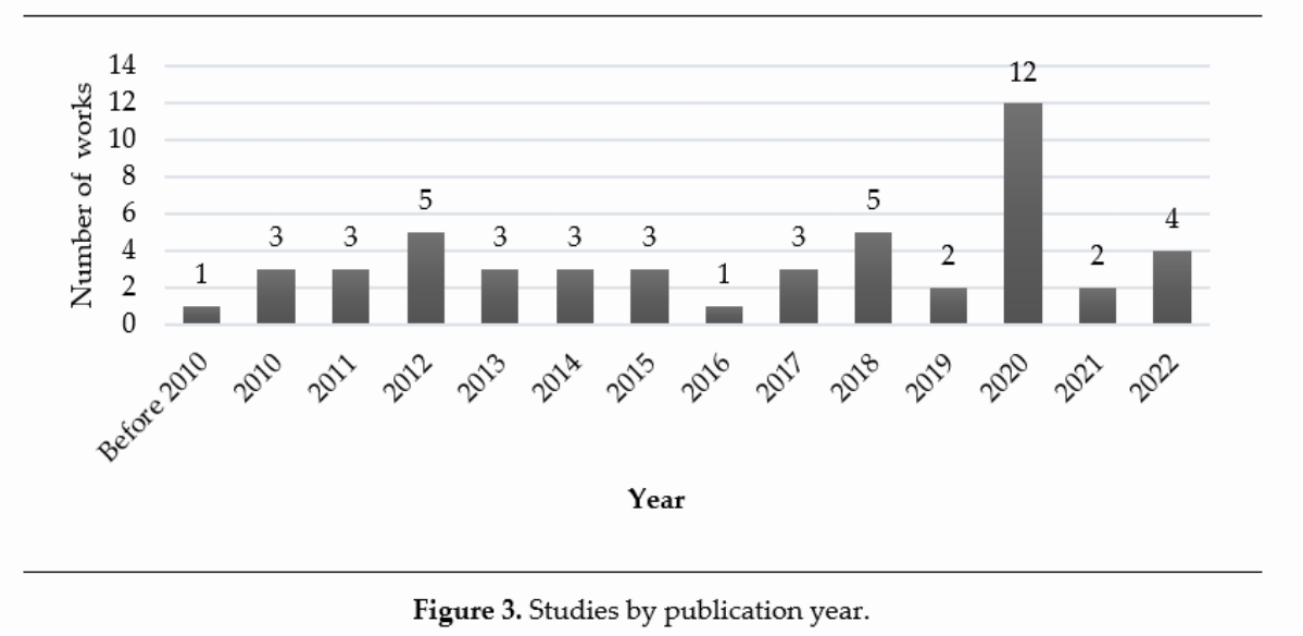

Figure 3. Studies by publication year.

3.2 Catalogue Summary Statistics: By Datasets (n = 40)

3.2.1 Types of Providers

We catalogued the RS datasets used in the curated publications, documenting their sources or providers. We categorized the sources into three groups: Government Agencies, Commercial Companies, and Academic Institutions (Table 2). The majority of the datasets (n = 28) were obtained from Government Agencies worldwide. Readers can refer to [10] for detailed information about the curated RS datasets and their corresponding providers and publications that featured them.

Published by Journal of the Academy of Public Health. This is an Open Access article under the CC BY 4.0 license.

Table 2. Dataset sources/providers by category (government, commercial, academic) and provider's region.

Provider Category	Provider Region	Providers	Number of Datasets
Government Agency	Asia	Earth Remote Sensing Data Analysis Center of Japan	1
		National Catalogue Service For Geographic Information of China	1
		National Geomatics Center of China	1
	Europe	European Space Agency (ESA)	5, including 1 dataset co-provided with Université Catholique de Louvain
		German Aerospace Center	1
		Ordnance Survey (UK)	3
		UK Centre for Ecology & Hydrology (UK CEH)	2
	America	Multi-Resolution Land Characteristics Consortium (USA)	1

Published by Journal of the Academy of Public Health. This is an Open Access article under the CC BY 4.0 license.

		National Aeronautics and Space Administration (NASA)	6, including one dataset co-provided with NOAA
		National Centers for Environmental Information, NESDIS, NOAA, U.S. Department of Commerce	2, including one dataset co-provided with NASA
		National Oceanic and Atmospheric Administration National Geophysical Data Center (USA)	1
		Panama's Ministry of the Environment (MiAmbiente; formerly ANAM)	1
		U.S. Geological Survey	2
		U.S. National Geospatial-Intelligence Agency (NGA)	1
		USDA's Farm Service Agency (FSA)	1
Commercial Company	America	Environmental Systems Research Institute (ESRI)	2
		Google	2
		Maxar Technologies	1
		Microsoft	1
	Europe	OpenStreetMap (Foundation)	1
		Vision On Technology	1

Published by Journal of the Academy of Public Health. This is an Open Access article under the CC BY 4.0 license.

Academic Institute	Asia	Data Center for Resources and Environmental Sciences of the Chinese Academy of Science	1
		Star Cloud Data Service Platform, Pengcheng Laboratory	1
	Europe	Copernicus Open Access Hub	1
		Université Catholique de Louvain	1 dataset co-provided with European Space Agency (ESA)
	America	Natural Resource Spatial Analysis Laboratory, Institute of Ecology, University of Georgia	1
Total			40 (co-provided datasets are only counted once)

3.2.2 Geographic Coverage

Table 3 presents the geographic coverage of the 40 catalogued datasets in our proof-of-concept catalogue. A considerable portion (about 60%) of these datasets encompasses a global scale, while the majority of the remaining datasets are available at a national or regional level. There are more datasets available for the USA and the UK compared with other countries in our proof-of-concept catalogue. Additionally, it is worth noting that we have included a dataset specifically covering oceanic regions. Readers can refer to [10] for detailed information about the curated datasets and their geographic coverage.

Published by Journal of the Academy of Public Health. This is an Open Access article under the CC BY 4.0 license.

Table 3. Dataset coverage.

Dataset Region	Covered Areas	Number of Datasets
Global	Global	24
	Global Coastal Zone (Ocean)	1
Americas	North America	1
	Panama	2
	USA	4
Asia	China	2
Africa	Kenya	1
Europe	England	1
	UK	4
Total		40

3.2.3 Costs

We also examined (and recorded in a dedicated metadata field in our catalogue) the cost of acquiring each of the curated RS datasets. Out of the 40 RS datasets in our proof-of-concept catalogue, 33 are freely accessible (free of charge), while the remaining seven require a purchase in order to use them for research purposes. Readers can refer to [10] for detailed information about the curated datasets and their cost status.

3.2.4 First Available Year and Update Frequency

Of the 40 datasets included in our proof-of-concept catalogue, 14 datasets were initially made available before the year 2000. A further 17 datasets were first released between 2000-2010, while the 9 remaining datasets became available for the first time after 2010. At least 21 out of the 40 curated datasets (52.5%) are still being updated as of 2022/2023 at various (usually regular) intervals since their initial release. The recorded update frequencies for these datasets range from daily updates and updates every ten days to monthly and quarterly updates to updates every five to ten years. Readers can refer to [10] for detailed information about the curated RS datasets and their corresponding first available year and update frequency details (where applicable and available).

3.2.5 Resolutions

Resolution is one of the critical factors in the selection of RS datasets as it needs to match the corresponding requirements of the application at hand. Table 5 provides details of the various dataset resolutions in our proof-of-concept catalogue. The most common dataset resolutions in the catalogue are 1km (7 datasets) and 30m (a further 7 datasets). Commercial companies such as Google Maps and Bing Maps provide RS datasets with notable resolution

Published by Journal of the Academy of Public Health. This is an Open Access article under the CC BY 4.0 license.

advantages (15cm–15m). It is important to note that some datasets offer multiple resolutions, as they contain remote sensing images captured in various bands. In such cases, panchromatic images generally have higher resolutions compared to multispectral images. Readers can refer to [10] for detailed information about the curated RS datasets and their various resolutions.

Table 5. Dataset resolutions (where available and recorded in our proof-of-concept catalogue). Some datasets offer multiple resolutions and therefore had to be counted more than once in this table.

Resolution Type	Resolution	Number of Datasets
Length	>10km	1
	1km	7
	800m	1
	300m	1
	250m	2
	84m	1
	30m	7
	20m	4
	10m	2
	5m	1
	2.4m / Multispectral	2
	0.6m / Panchromatic	3, including the same two multispectral datasets counted in above row, as they are also panchromatic
	0.15–0.5m	4
	N/A	2
Scale	1:10,000	1
	1:100,000	1
	1:250,000	1
	1:7,000,000	1

3.3 GeoExposomics.org Online Portal and Searchable Catalogue

The project's portal offers a searchable version of the catalogue, and can be accessed at [11][11] (Figures 2 and 3). It allows users to search and filter dataset records by health conditions/application areas, cost, covered area(s), provider, or a combination of one or more of these elements. Users can also filter the associated research article records by health conditions/application areas and/or dataset name.

Published by Journal of the Academy of Public Health. This is an Open Access article under the CC BY 4.0 license.

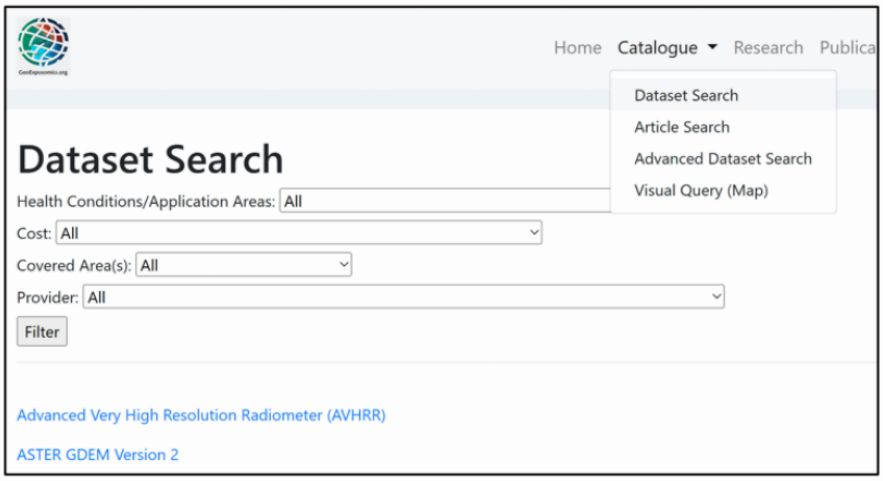

Figure 2. An online searchable version of the catalogue can be accessed by clicking the 'Catalogue' menu on the project portal's navigation bar. Various search modes and result filtering options are currently available.

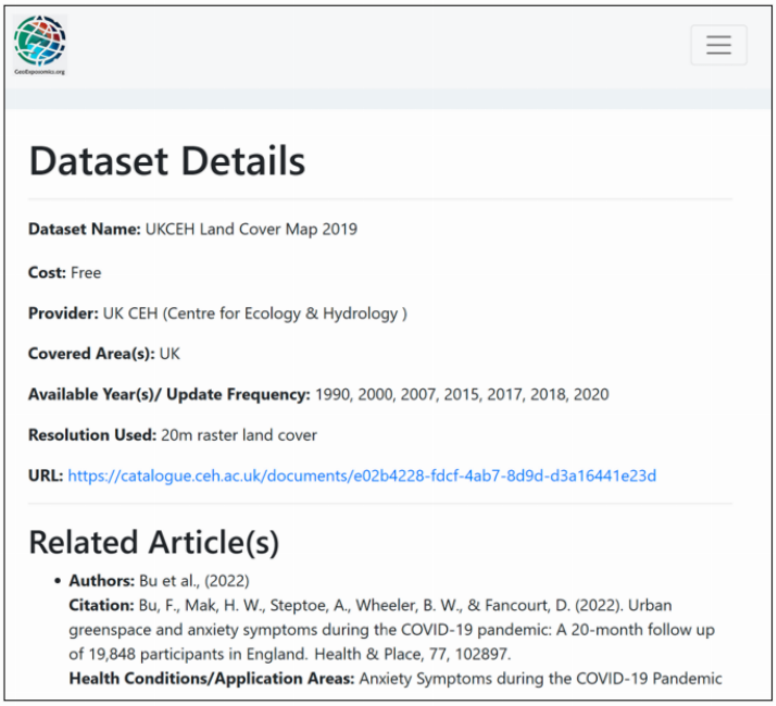
Dataset Details
Dataset Name: UKCEH Land Cover Map 2019
Cost: Free
Provider: UK CEH (Centre for Ecology & Hydrology)
Covered Area(s): UK
Available Year(s)/ Update Frequency: 1990, 2000, 2007, 2015, 2017, 2018, 2020
Resolution Used: 20m raster land cover
URL: https://catalogue.ceh.ac.uk/documents/e02b4228-fdcf-4ab7-8d9d-d3a16441e23d
Related Article(s)
Authors: Bu et al., (2022)
Citation: Bu, F., Mak, H. W., Steptoe, A., Wheeler, B. W., & Fancourt, D. (2022). Urban greenspace and anxiety symptoms during the COVID-19 pandemic: A 20-month follow up of 19,848 participants in England. Health & Place, 77, 102897.
Health Conditions/Application Areas: Anxiety Symptoms during the COVID-19 Pandemic

Figure 3. A typical metadata record in the online catalogue. This screenshot shows the mobile version of the interface with the 'hamburger menu' (top right) for site navigation.

3.4 Evaluation Feedback and Second Iteration Improvements

The evaluation feedback received from our panel of five stakeholder representatives about the online portal was positive and provided a couple of valuable suggestions that were later implemented in a second iteration of the interface (released in late June 2023), namely a suggestion "to implement free text search and ensure it can tolerate different spelling variants (e.g., British and American) and misspellings of health applications and place names" and another one to have a handy "online map to help locating datasets". In response, we implemented a free text search option of datasets (by health applications or covered areas) using a custom Soundex algorithm[12] (Figure 4) and introduced a visual query option using a map with clickable hotspots (Figure 5). It should be noted that while free text search using Soundex is an improvement, it still cannot help when a query string is a synonym, i.e., a very different word or phrase not in the catalogue as such; for example, when searching for 'coronavirus disease' instead of 'COVID-19'. This remains a limitation of the current search interface.

Published by Journal of the Academy of Public Health. This is an Open Access article under the CC BY 4.0 license.

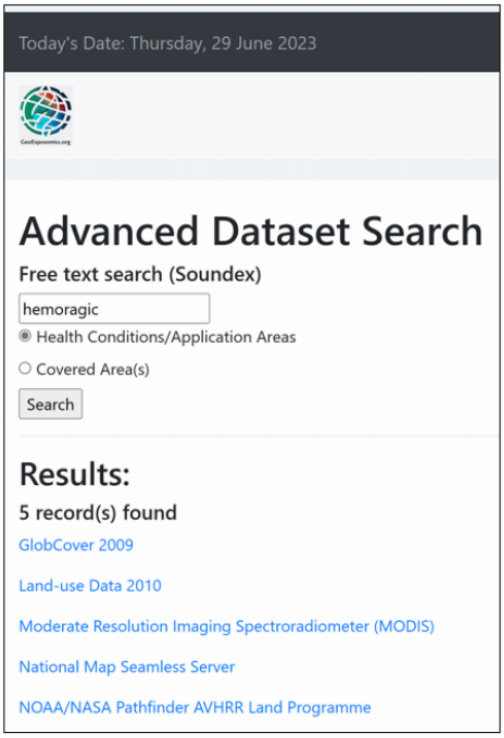

Figure 4. Following user evaluation feedback, we implemented a free text search option of datasets (by health applications or covered areas) using a custom Soundex algorithm. As an example of how this algorithm works, searching health conditions for 'haemorrhagic fever' (British spelling), 'hemorrhagic fever' (American spelling), or 'hemoragic fever' (typo/misspelling, as in the above screenshot) will all retrieve the correct/same results.

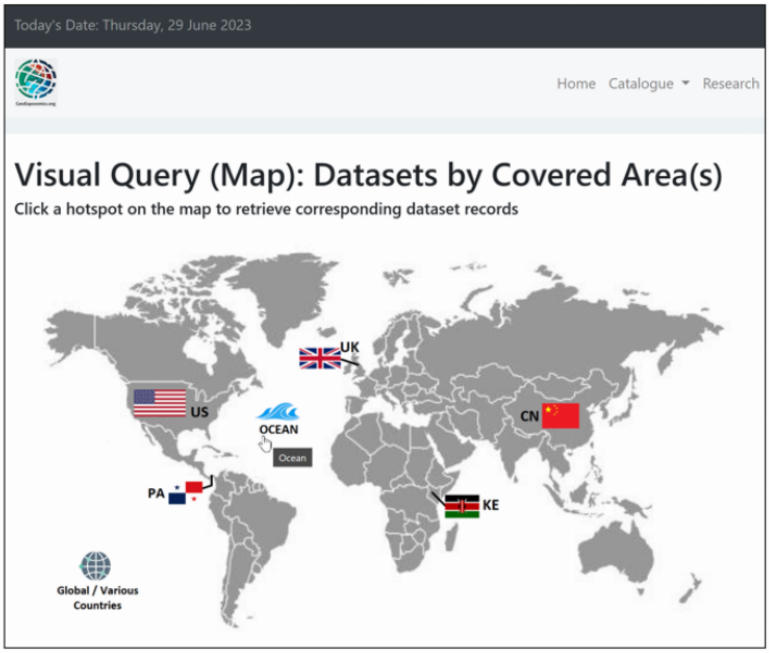

Figure 5. User evaluation feedback included a suggestion to provide an online map to visually locate datasets. In response, we introduced an option allowing users to click location 'hotspots' on a map of the world with country flags and other relevant icons to retrieve corresponding dataset records.

3.5 Dissemination

The portal was widely advertised on social media upon its initial launch in June 2023 and subsequent updates during June and July 2023 using the P.I.'s (MN Kamel Boulos) accounts on Facebook, Instagram, LinkedIn, and Twitter (>2700 combined followers across the four platforms), with additional reposts via other relevant public LinkedIn and Twitter accounts. Between 15 June and 31 July 2023, the portal received >550 unique visitors from more than 25 countries according to anonymous server log statistics generated by AWStats tool, an open-source server log file analyzer for

Published by Journal of the Academy of Public Health. This is an Open Access article under the CC BY 4.0 license.

advanced statistics, using site visitors' IP geolocation. We also have a companion Facebook page where relevant news items and updates are regularly shared with a growing community.[13]

4 - Discussion

RS techniques can help us better map and investigate the exposomic determinants of disease (population-level exposures) by providing spatially and temporally resolved data over time at multiple geographic scales. For example, satellite imagery can be used to monitor air quality and temperature, and detect the presence of pollutants or heat waves and extreme weather events that may have adverse effects on human health.[14,15,16] RS data can also aid in identifying patterns of land use and land cover that contribute to the spread of diseases such as malaria[17] and dengue fever.[18] However, discovering and locating relevant RS datasets for use in a particular health study can often prove challenging, hence the need for metadata catalogue services to assist researchers in this task.

4.1 Metadata Catalogue Services

Catalogue services, such as our proof-of-concept portal [11], allow for the publishing and searching of collections of descriptive information (or metadata) about resources: data, services, and related information objects, or in our case, metadata about RS datasets of interest. As such, metadata catalogue services are essential for supporting "the discovery and binding to registered information resources within an information community,"[19] which in the case of our project would be the 'RS and population health and medicine research communities'. It is worth noting here that the European Union's INSPIRE (Infrastructure for Spatial Information in Europe) portal in partnership with UNIGIS offer a highly relevant online training module entitled 'metadata and catalogue services' and aimed at data users and data providers.[20]

4.2 Reflections on Our Catalogue Summary Statistics

In our analysis of the publications included in our proof-of-concept catalogue, we observed that more studies were published in geography and environmental science journals than in public health journals. However, with our very small convenience sample of five experts and 50 curated publications, one cannot boldly generalize this observation. Nevertheless, there are a couple of reasons as to why RS data are still relatively underused in population health studies.

Firstly, some researchers with a primary background in health might be unfamiliar with RS data and their potential applications in health studies. Secondly, in order to use RS data proficiently, researchers should also be familiar with, or seek interdisciplinary collaborations with geomatics researchers who are familiar with, the different types and formats of RS data. In general, RS data come from imagery. To digitally manage and process imagery data, RS data are constructed in the form of pixels (cells) or in raster, the rectangular arrays of numbers containing both geometric and radiometric characteristics from a surveyed area, often extended with other attribute data. For efficient data management and analysis, raster data are written and managed in several different file formats to encode a variety of geographic data structures. Therefore, conceptually understanding and utilizing RS data in raster format may be challenging for health researchers who are usually more familiar with numbers-only data written in a simpler structure.

Another observation relates to the study areas of the catalogued publications and their corresponding dataset coverage. Only a handful of studies in our sample examined problems in the Global South. Again, this observation cannot be readily generalized given our small sample size, but it still has some merit and is worth considering. Unarguably, there are many public health and environmental challenges to address in the Global North, especially in light of climate change [6] and population ageing. However, more research should be conducted to investigate such challenges in the Global South, considering (and addressing) the uneven development in research experience, capacity, and infrastructure there.

4.2.1 Scale and Resolution Matter

Thanks to rapid technological developments, it is becoming increasingly easier nowadays to use RS data with low, medium, and high spatial resolution. A commercial satellite offers a spatial resolution of 30cm x 30cm, which is about the finest spatial resolution for now. It is not always the case, but a higher spatial resolution could be helpful in a closer examination of the impact of the environment on population health as it can provide finer details. However, obtaining higher resolution RS data is costly, and there is not always a clear consensus about what level of resolution

Published by Journal of the Academy of Public Health. This is an Open Access article under the CC BY 4.0 license.

would be optimal for use in a given health application or study. Considering the temporal resolution, i.e., the frequency of the revisit period, could be another challenge when controlling for the continuing interaction between humans and the environment.

4.2.2 The Importance of Data Linkages

The majority of our catalogued RS datasets are publicly available free of charge. Two of the most notable and frequently cited RS data portals in our proof-of-concept catalogue are NASA's Earthdata[21] and ESA's Sentinel Online.[22] These data portals provide researchers and any interested members of the public with one-stop services covering a range of user needs and topics from RS data access to various research themes and case studies to self-learning material. However, these data portals only serve as a platform for RS data and related information, without offering any linkages to population-related data. In contrast, the applications of GIS (Geographic Information Systems) have become standard in the public health literature, with the US Centers for Disease Control and Prevention (CDC) now operating a database service linking relevant demographic, social, and environmental data, as well as GIS software and tools.[23] Future development in RS data portals should explore potential ways of linking RS data to other datasets to enable comprehensive research in disciplines requiring such linkages, such as population health and medicine.

4.3 Future Directions: Turning Current Limitations into Opportunities for Future Improvements

Our current small sample size and online user interface limitations (e.g., free text search not catering for synonyms) can be used to inform, and set the desiderata for, possible future developments. A follow-on project would conduct a much wider stakeholder consultation and comprehensively catalogue more datasets covering more countries/locations and more health conditions/application areas. Moreover, an online form could be made available on the portal for visitors to propose and contribute metadata details of new records to be added to the Catalogue (if deemed suitable following moderation to ensure quality and completeness).

As the catalogue expands, revisiting and improving our metadata schema, elements and descriptions' consistency and organization, expanding them as appropriate, and applying the relevant standards of metadata and catalogue services[24] [19,20,24] will become imperative. A protocol with the necessary mechanisms for catalogue ongoing maintenance will need to be established to ensure the continued currency of the curated metadata records and the details stored within.

We could also look into providing a ChatGPT-like,[25,26,27,28] natural language-based AI (artificial intelligence)-driven conversational user interface (UI) for querying the catalogue, whereby users use free text (plus, optionally, a more advanced interactive map to select desired locations) to describe in their own way and words a research question or health application they are currently investigating in a particular locale in order to retrieve all relevant records from the catalogue. This would enhance discovery and explorability by allowing the search UI to more flexibly adapt to users' needs in more natural ways instead of asking users to adapt to and master a rather rigid (preprogrammed) and less forgiving interface.

The implementation of a ChatGPT-like interface will make use of Retrieval-Augmented Generation (RAG) whereby our metadata catalogue becomes an external knowledgebase to an existing Large Language Model (LLM), such as GPT-4, in order to provide more relevant and better cross-linked search results in response to user's prompts[29] (Figure 6).

Published by Journal of the Academy of Public Health. This is an Open Access article under the CC BY 4.0 license.

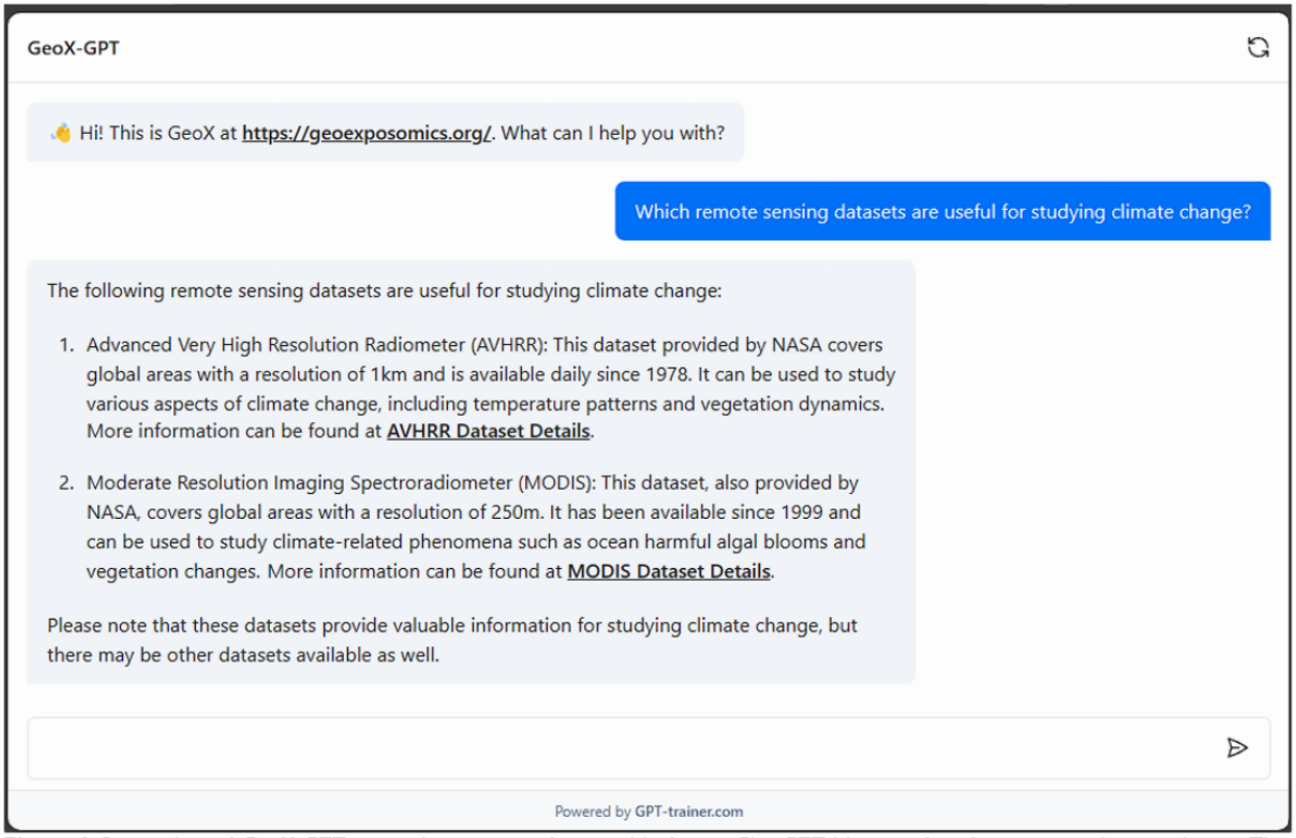
GeoX-GPT
Hi! This is GeoX at https://geoexposomics.org/. What can I help you with?
Which remote sensing datasets are useful for studying climate change?
The following remote sensing datasets are useful for studying climate change:
1. Advanced Very High Resolution Radiometer (AVHRR): This dataset provided by NASA covers global areas with a resolution of 1km and is available daily since 1978. It can be used to study various aspects of climate change, including temperature patterns and vegetation dynamics. More information can be found at AVHRR Dataset Details.
2. Moderate Resolution Imaging Spectroradiometer (MODIS): This dataset, also provided by NASA, covers global areas with a resolution of 250m. It has been available since 1999 and can be used to study climate-related phenomena such as ocean harmful algal blooms and vegetation changes. More information can be found at MODIS Dataset Details.
Please note that these datasets provide valuable information for studying climate change, but there may be other datasets available as well.
Powered by GPT-trainer.com

Figure 6. Screenshot of GeoX-GPT, an early preview of a possible future ChatGPT-like interface for our metadata catalogue. This experimental RAG implementation was developed in GPT-trainer [30].

Figure 6. Screenshot of GeoX-GPT, an early preview of a possible future ChatGPT-like interface for our metadata catalogue. This experimental RAG implementation was developed in GPT-trainer.[30]

5 - Conclusions

Earth observation data play an important role in understanding the relationship between environmental factors and human health, offering valuable insights into the exposomic determinants of disease (population-level exposures). During 2023, the GeoExposomics Project Team built and published a proof-of-concept online RS data catalogue to further facilitate human health research in exposomics, through an international online collaboration between the Project Team members themselves (who are based in five countries spanning four continents) and with an external panel of five leading scholars from three countries.

In this article, we conducted an analysis of the content of our proof-of-concept catalogue and described the steps taken to convert it into a user-friendly online searchable format over two development iterations informed by user feedback. We also provided some reflections on our catalogue summary statistics, our current project limitations, and possible future developments. It is hoped future, more comprehensive versions of this service will offer a complete one-stop shop for discovering health-related RS datasets from multiple providers worldwide, enabling more interdisciplinary researchers and human exposomics studies to discover and use Earth observation data in combination with other relevant data to reveal fresh insights that could improve our understanding of relevant diseases, and hence contribute to the development of better-optimized prevention and management strategies to tackle them.

6 - Important Note to Readers

This work is strictly a proof of concept (PoC). PoCs have inherent major limitations compared with more developed prototypes and pilots. The latter would have required much larger budgets and resources to execute than what we had at our disposal for this project. For a detailed comparison of PoCs, prototypes and pilots, please consult [31].[31]

Published by Journal of the Academy of Public Health. This is an Open Access article under the CC BY 4.0 license.

Author Contributions: Conceptualization, M.N.K.B.; methodology, M.N.K.B. (Main) and K.K.; validation, H.Z. and M.N.K.B.; formal analysis and investigation, K.K., G.Z. and M.N.K.B.; resources, M.N.K.B.; data curation, K.K. (Main), G.Z. (Main) and M.N.K.B.; writing—original draft preparation, K.K., G.Z. and M.N.K.B.; writing—review & editing, K.K., G.Z., M.N.K.B., H.Z., M.V.I, B.B. and A.D.; visualization, K.K., G.Z. and M.N.K.B.; supervision and project administration, M.N.K.B. (Main/Lead), K.K., M.V.I, B.B. and A.D.; funding acquisition, M.N.K.B. (P.I.) with M.V.I, B.B. and A.D. (Co-Is). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the International Society for Photogrammetry and Remote Sensing (ISPRS) under ISPRS Scientific Initiatives 2023 (SI2023 Awards).

Data Availability Statement: The proof-of-concept catalogue data covered in this paper are publicly and freely available from doi: 10.5281/zenodo.8186444 and https://geoexposomics.org/

Acknowledgments: The authors would like to express their gratitude to the panel of stakeholder representatives that were involved in the project.

Conflicts of Interest: The authors declare no conflict of interest.

Published by Journal of the Academy of Public Health. This is an Open Access article under the CC BY 4.0 license.

References

1.

Toth C, Jóźków G. Remote sensing platforms and sensors: A survey. ISPRS Journal of Photogrammetry and Remote Sensing 2016, 115:22–36. doi: 10.1016/j.isprsjprs.2015.10.004.

2.

Chi M, Plaza A, Benediktsson JA, Sun Z, Shen J, Zhu Y. Big data for remote sensing: Challenges and opportunities. Proceedings of the IEEE 2016, 104(11):2207–2219. doi: 10.1109/JPROC.2016.2598228.

3.

The US National Institute for Occupational Safety and Health (NIOSH). Exposome and Exposomics. https://www.cdc.gov/niosh/topics/exposome/default.html (accessed Nov 14, 2022).

4.

Cui Y, Eccles KM, Kwok RK, Joubert BR, Messier KP, Balshaw DM. Integrating Multiscale Geospatial Environmental Data into Large Population Health Studies: Challenges and Opportunities. Toxics 2022, 10(7):403. doi: 10.3390/toxics10070403.

5.

Kamel Boulos MN, Le Blond J. On the road to personalised and precision geomedicine: medical geology and a renewed call for interdisciplinarity. Int J Health Geogr 2016, 15(5). doi: 10.1186/s12942-016-0033-0.

6.

Kamel Boulos MN, Wilson JP. Geospatial techniques for monitoring and mitigating climate change and its effects on human health. Int J Health Geogr 2023, 22(2). doi: 10.1186/s12942-023-00324-9.

7.

Sogno P, Traidl-Hoffmann C, Kuenzer C. Earth Observation Data Supporting Non-Communicable Disease Research: A Review. Remote Sensing 2020, 12(16):2541. doi: 10.3390/rs12162541.

8.

Urbano T, Vinceti M, Wise LA, Filippini T. Light at night and risk of breast cancer: a systematic review and dose–response meta-analysis. Int J Health Geogr 2021, 20(44). doi: 10.1186/s12942-021-00297-7.

9.

International Society for Photogrammetry and Remote Sensing. ISPRS Announcements: Seven Projects win ISPRS Scientific Initiatives 2023 (SI2023) Awards. https://www.isprs.org/news/announcements/details.aspx?ID=314 (accessed Feb 27, 2023).

10.

Kamel Boulos MN, GeoExposomics Project Team. GeoExposomics.org proof-of-concept catalogue as an Excel workbook (Catalogue Main Version: 10 May 2023; Last Updated 25 July 2023) [Data set]. Zenodo 2023, doi: 10.5281/zenodo.8186444.

11.

Kamel Boulos MN (P.I.), GeoExposomics Project Team. GeoExposomics.org Portal and Online Catalogue. https://geoexposomics.org/ (accessed Jul 24, 2023).

Published by Journal of the Academy of Public Health. This is an Open Access article under the CC BY 4.0 license.

12.

US National Archives. Soundex System. https://www.archives.gov/research/census/soundex (accessed Jun 16, 2023).

13.

Kamel Boulos MN (P.I.), GeoExposomics Project Team. GeoExposomics Facebook Page. https://www.facebook.com/geoexposomics (accessed Jul 24, 2023).

14.

Fuentes M, Millard K, Laurin E. Big geospatial data analysis for Canada's Air Pollutant Emissions Inventory (APEI): using Google Earth engine to estimate particulate matter from exposed mine disturbance areas. GIScience & Remote Sensing 2020, 57(2), 245-257. doi: 10.1080/15481603.2019.1695407.

15.

Guo H, Li W, Yao F, Wu J, Zhou X, Yue Y, Yeh AGO. Who are more exposed to PM2.5 pollution: A mobile phone data approach. Environ Int 2020, 143, 105821. doi: 10.1016/j.envint.2020.105821.

16.

Kloog I, Zhang X. Methods to Advance Climate Science in Respiratory Health: Satellite-Based Environmental Modeling for Temperature Exposure Assessment in Epidemiological Studies. Immunol Allergy Clin North Am 2024, 44(1), 97-107. doi: 10.1016/j.iac.2023.07.002.

17.

Solís P, McCusker B, Menkiti N, Cowan N, Blevins C. Engaging global youth in participatory spatial data creation for the UN sustainable development goals: The case of open mapping for malaria prevention. Applied Geography 2018, 98, 143-155. doi: 10.1016/j.apgeog.2018.07.013.

18.

Kamal ASMM, Al-Montakim MN, Hasan MA, Mitu MMP, Gazi MY, Uddin MM, Mia MB. Relationship between Urban Environmental Components and Dengue Prevalence in Dhaka City-An Approach of Spatial Analysis of Satellite Remote Sensing, Hydro-Climatic, and Census Dengue Data. Int J Environ Res Public Health 2023, 20(5), 3858. doi: 10.3390/ijerph20053858.

19.

Open Geospatial Consortium. Catalogue Service. https://www.ogc.org/standard/cat/ (accessed Jul 26, 2023).

20.

EU INSPIRE and UNIGIS. Metadata and Catalogue Services (training module). https://inspire.ec.europa.eu/training/metadata-and-catalogue-services (accessed Jul 26, 2023).

21.

US National Aeronautics and Space Administration. NASA Earthdata. https://www.earthdata.nasa.gov/ (accessed Jul 26, 2023).

22.

European Space Agency. Sentinel Online. https://sentinels.copernicus.eu/web/sentinel/home (accessed Jul 26, 2023).

23.

US Centers for Disease Control and Prevention. GIS and Public Health at CDC: Geospatial Data Resources. https://www.cdc.gov/gis/geo-spatial-data.html (accessed Jul 26, 2023).

Published by Journal of the Academy of Public Health. This is an Open Access article under the CC BY 4.0 license.

24.

Mathys T, Kamel Boulos MN. Geospatial resources for supporting data standards, guidance and best practice in health informatics. BMC Res Notes 2011, 4, 19. doi: 10.1186/1756-0500-4-19.

25.

Tao R, Xu J. Mapping with ChatGPT. ISPRS International Journal of Geo-Information 2023, 12(7), 284. doi: 10.3390/ijgi12070284.

26.

Gautam R. OSM-GPT: an innovative project combining GPT-3 and the Overpass API to facilitate easy feature discovery on OpenStreetMap. https://github.com/rowheat02/osm-gpt (accessed Jul 26, 2023).

27.

Velvárt A. How will AI Affect User Interfaces? Linkedin 2023. https://www.linkedin.com/pulse/how-ai-affect-user-interfaces-andr%2525C3%2525A1s-velv%2525C3%2525A1rt/ (accessed Jul 26, 2023).

28.

Vesal M. The Evolution of UX Design and the Impact of ChatGPT and Other AI Technologies. Linkedin 2023. https://www.linkedin.com/pulse/evolution-ux-design-impact-chatgpt-other-ai-majid-vesal/ (accessed Jul 26, 2023).

29.

Open AI. Retrieval Augmented Generation (RAG) and Semantic Search for GPTs. https://help.openai.com/en/articles/8868588-retrieval-augmented-generation-rag-and-semantic-search-for-gpts (accessed 19 March 2024)

30.

Petal/Paladin Max, Inc. GPT-trainer. https://gpt-trainer.com/ (accessed on 29 March 2024).

31.

The London School of Economics and Political Science (LSE). Use of Proof of Concepts, Prototypes and Pilots. https://info.lse.ac.uk/staff/services/Policies-and-procedures/Assets/Documents/proConProPil.pdf (accessed on 29 March 2024).

Published by Journal of the Academy of Public Health. This is an Open Access article under the CC BY 4.0 license.

Peer Review of: A Proof-of-Concept Metadata Catalogue and Online Portal of Earth Observation Datasets for Health Research in Exposomics

Gary Shaw

Stanford University School of Medicine

DOI: https://doi.org/10.70542/rcj-japh-pr-166alts

Koh and colleagues have written a very nice narrative describing their proof-of-concept development of a metadata catalog that characterizes remote-sensing data available to researchers. Their objective is clear and their efforts much appreciated. As they note, use of such data in population health investigations is not a new endeavor but having such data more easily within the reach of far more investigators than the mere few in-the-know is critical to advancing population health investigations. As an epidemiologist who focuses on exposomics and perinatal outcomes, I certainly welcome such a nice annotation/curation of available earth-observational data.

Koh and colleagues effectively summarize key aspects of the catalog and transparently acknowledge some limitations of their proof-of-concept work. They might consider adding to their narrative (albeit an obvious caveat) that various remote sensing datasets may have highly variable accuracy and thus user-beware is potentially worth mentioning to those readers less experienced in exposure measurement. Future developments that expand on this proof-of-concept will be an important resource to researchers in the manifold fields that comprise population health science.

Learning from the catalog that a data investigation from one region or time period could be potentially replicated in another region and/or time is an essential ingredient toward furthering the rigor and reproducibility of research writ large. As the authors suggest, the platform that has been built could chronicle efforts (investigators or their publications) that used a particular data type/set for one disease to guide researchers through the strengths/limitations who are investigating entirely different diseases. This cataloging endeavor also holds the potential to foster global collaborative research opportunities.

Peer Review of: A Proof-of-Concept Metadata Catalogue and Online Portal of Earth Observation Datasets for Health Research in Exposomics

Ivair Silva

Federal University of Ouro Preto

DOI: https://doi.org/10.70542/rcj-japh-pr-9yb4yj

ivair@ufop.edu.br

Overview

The manuscript discusses the role of metadata catalogues and observational online data in enhancing exposome studies, helping to prevent and manage exposomic determinants of disease worldwide. For this, the authors report the results of recent studies using free information from the International Society for Photogrammetry and Remote Sensing (ISPRS)-funded project, and a proof-of-concept catalogue provided by desk research involving 90 metadata records, 40 Earth observation datasets, and 50 research publications. The content of this paper is interesting and scientifically relevant to several fields.

This review follows the plus-delta-interpretation format. Each of these three aspects are described separately in

Published by Journal of the Academy of Public Health. This is an Open Access article under the CC BY 4.0 license.

what follows.

Plus

1. The paper contributes to articulating the importance of Earth observation data for research in population health and medicine. Because the exposomics deals with the interaction of multiple factors that can lead to the development of different types of diseases (communicable and non- communicable), the discussion of potential mechanisms to link the multiple factors (genomic, environmental, lifestyle) is deterministic for the comprehensive understanding, prevention and management of potential diseases that can threaten the public health.

2. The paper is well-written and friendly structured. The ideas and concepts are introduced and explained in such a way that readers from virtually any scientific field can fully comprehend the importance of Earth observation data sets for public health.

3. The existence of an online portal metadata catalogue is very important for this field. The dissemination of the portal and its features is one of the major contributions from this manuscript.

4. Although modest in number, the 50 research publications and the 5 stakeholders provided rich information to understand the potential contribution of remote sensing (RS) data to the identification of causes of diseases and mitigation of related potential risks.

Delta

1. The goals of the paper can be inferred with the explanations in the Abstract and in the second paragraph of Section 1.2. Still, I feel that the main objective could be stated more directly. I understand that this manuscript aims to articulate the importance of Earth observation data sources/products and types for the reseach in health related fields. On the other hand, in some parts of the text it appeared to me that the main goal is to highlight the importance of the Catalogue and the online Portal. Hence, stating the main objectives of the paper can help to locate and link such important aspects presented in the text.

2. The Introduction could contain more information about the ISPRS, e.g the year of foundation, the location of main operation, funders (is it a governement agency, private company, or a non-profit society?), and a little history about this initiative.

3. Some key results of this work could be succinctly mentioned by the end of Section 1.2. This would help readers to better visualize the goals of the paper. For example, the authors could anticipate the scale and resolution of geographic information matters, the importance of data linkages and its free availiability, and the opportunities for future improvements.

4. What about mentioning whether the five stakeholder representatives in the desk research were asked about possible conflicts of interest (COI) they may have related to the available products in public health that they mentioned? Naturally, the existence of financial or non-financial COI does not disable their opinion. The information on COI benefits transparency. If this information was not asked to the five experts, then this could be stated in the paper.

5. I would suggest the authors to include a brief overview on other local or international initiatives, similar to ISPRS, describing the diferences and the features they have in common with ISPRS. For example, I wonder if there exist similar catalogues and online portals developed by other institutions around the world. The discussion of other initiatives is another way to highlight and facilitate the use of Earth observation data for furtherresearch in exposomics.

6. A simple example with some RS data could be offered, such as a print of the data in an Excel file. This could help researchers to better understand how these data appear in practice and how they can be associated, under a exposomic approach, as a determinant of disease.

Interpretation

The project described in this paper seems to be a premier initiative for the development of tools for Earth observation data and for the creation of a metadata catalogue.

Published by Journal of the Academy of Public Health. This is an Open Access article under the CC BY 4.0 license.

One remarkable characteristic that can be interpreted from the results of this study is that this subject has been explored mainly by developed countries. This is evidenced with the mapping of dataset source/provider regions (Table 2). It is surprising that there are no other countries among the provider agencies in the American continent other than the United States. The fact that only three countries were represented in the desk research is also informative about the lack of research from most of the countries around the world on using Earth observation data.

This paper also clarifies that metadata catalogues are crucial to discover and analyze RS information and connect it with the public health and medical research communities.

Apparently, it is hard to obtain volunteers to compose the desk research. The authors' plans for future improvements can potentially alleviate this problem in order to embrace more countries and health application areas.

Author Rejoinder

Maged N. Kamel Boulos

School of Medicine, University of Lisbon

mnkboulos@ieee.org

We would like to thank both referees for their valuable and informative comments.

As the first referee (Dr. Gary Shaw) rightly noted, "various remote sensing datasets may have highly variable accuracy." Users who are less experienced in exposure measurement would definitely benefit from this tip. In fact, one possible future direction for this project would be to expand our experimental LLM (Large Language Model) and RAG (Retrieval-Augmented Generation)-based GeoX-GPT conversational interface (shown in Figure 6 in the paper) and its current purpose, which is limited to querying the metadata catalogue, into a fully-fledged agentic system (AI agent) that contextually incorporates and automates this tip and others (the 'know-how' of expert users) to assist less experienced users in selecting, preparing, and running the most appropriate datasets, approaches and methods for a given investigation or research question. (Please refer to Google's White Paper "Agents" by Wiesinger et al., 2024 at https://www.kaggle.com/whitepaper-agents and Anthropic's December 2024 article entitled "Building effective agents" at https://www.anthropic.com/research/building-effective-agents for more information about agentic systems.)

The second referee (Dr. Ivair Silva) rightly captured the two-pronged goals of this research, namely (1) to highlight the importance of, and curate, a proof-of-concept metadata catalogue of Earth observation data sources, products and types that are relevant to human health research in exposomics, and (2) to examine a number of complementary user interface and experience options to make the catalogue searchable and more accessible to its end users, and to improve the discoverability of its content (the online portal).

Established in 1910, the International Society for Photogrammetry and Remote Sensing (ISPRS - https://www.isprs.org/) is an international non-governmental organization. In 2014, ISPRS introduced its Scientific Initiatives grant programme to support projects of interest to its community (https://www.isprs.org/society/si/default.aspx). Our project was funded by ISPRS under Scientific Initiatives 2023.

Published by Journal of the Academy of Public Health. This is an Open Access article under the CC BY 4.0 license.